Working Locally Thinking Globally: Theoretical Guarantees for Convolutional Sparse Coding

Vardan Papyan, Jeremias Sulam, *Student Member, IEEE*, and Michael Elad, *Fellow, IEEE*

***Abstract*—The celebrated sparse representation model has led to remarkable results in various signal processing tasks in the last decade. However, despite its initial purpose of serving as a global prior for entire signals, it has been commonly used for modeling low dimensional patches due to the computational constraints it entails when deployed with learned dictionaries. A way around this problem has been recently proposed, adopting a convolutional sparse representation model. This approach assumes that the global dictionary is a concatenation of banded circulant matrices. While several works have presented algorithmic solutions to the global pursuit problem under this new model, very few truly-effective guarantees are known for the success of such methods. In this paper, we address the theoretical aspects of the convolutional sparse model providing the first meaningful answers to questions of uniqueness of solutions and success of pursuit algorithms, both greedy and convex relaxations, in ideal and noisy regimes. To this end, we generalize mathematical quantities, such as the ℓ_0 norm, mutual coherence, Spark and restricted isometry property to their counterparts in the convolutional setting, intrinsically capturing local measures of the global model. On the algorithmic side, we demonstrate how to solve the global pursuit problem by using simple local processing, thus offering a first of its kind bridge between global modeling of signals and their patch-based local treatment.**

***Index Terms*—Sparse representations, convolutional sparse coding, uniqueness guarantees, stability guarantees, orthogonal matching pursuit, basis pursuit, global modeling, local processing.**

I. Introduction

A POPULAR choice for a signal model, which has proven to be very effective in a wide range of applications, is the celebrated sparse representation prior [1]–[4]. In this framework, one assumes a signal $\mathbf{X} \in \mathbb{R}^N$ to be a sparse combination of a few columns (atoms) $\mathbf{d}_i$ from a collection $\mathbf{D} \in \mathbb{R}^{N \times M}$, termed dictionary. In other words, $\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} \in \mathbb{R}^M$ is a sparse vector. Finding such a vector can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\Gamma}} g(\boldsymbol{\Gamma}) \ \text{ s.t. } \mathbf{D}\boldsymbol{\Gamma} = \mathbf{X}, \tag{1}$$

where $g(\cdot)$ is a function which penalizes dense solutions, such as the ℓ_1 or ℓ_0 "norms".[1] For many years, analytically defined matrices or operators were used as the dictionary $\mathbf{D}$ [5], [6]. However, designing a model from real examples by some learning procedure has proven to be more effective, providing sparser solutions [7]–[9]. This led to vast work that deploys dictionary learning in a variety of applications [4], [10]–[13].

Generally, solving a pursuit problem is a computationally challenging task. As a consequence, most such recent successful methods have been deployed on relatively small dimensional signals, commonly referred to as *patches*. Under this *local* paradigm, the signal is broken into overlapped blocks and the above defined sparse coding problem is reformulated as

$$\forall\, i \quad \min_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) \ \text{ s.t. } \mathbf{D}_L\boldsymbol{\alpha} = \mathbf{R}_i\mathbf{X},$$

where $\mathbf{D}_L \in \mathbb{R}^{n \times m}$ is a local dictionary, and $\mathbf{R}_i \in \mathbb{R}^{n \times N}$ is an operator which extracts a small local patch of length $n \ll N$ from the global signal $\mathbf{X}$. In this set-up, one processes each patch independently and then aggregates the estimated results using plain averaging in order to recover the global reconstructed signal. A local-global gap naturally arises when solving global tasks with this local approach, which ignores the correlation between overlapping patches. The reader is referred to [14]–[19] for further insights on this dichotomy.

The above discussion suggests that in order to find a consistent global representation for the signal, one should propose a global sparse model. However, employing a general global (unconstrained) dictionary is infeasible due the computational complexity involved, and training this model suffers from the curse of dimensionality. An alternative is a (constrained) global model in which the signal is composed as a superposition of local atoms. The family of dictionaries giving rise to such signals is a concatenation of banded Circulant matrices. This global model benefits from having a local shift invariant structure – a popular assumption in signal and image processing – suggesting an interesting connection to the above-mentioned local modeling.

When the dictionary $\mathbf{D}$ has this structure of a concatenation of banded Circulant matrices, the pursuit problem in (1) is usually

Manuscript received April 7, 2017; accepted July 17, 2017. Date of publication July 31, 2017; date of current version August 31, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gesualdo Scutari. This work was supported in part by the European Research Council under EU's seventh Framework Program, ERC under Grant 320649, and in part by Israel Science Foundation under Grant 1770/14. *(Vardan Papyan and Jeremias Sulam contributed equally to this work.) (Corresponding author: Vardan Papyan.)*

The authors are with the Computer Science Department, Technion - Israel Institute of Technology, Haifa 3200003, Israel (e-mail: vardanp91@gmail.com; jsulam@cs.technion.ac.il; elad@cs.technion.ac.il).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org. This material contains proofs of the theorems presented in the manuscript. The file is 698 KB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2733447

[1]Despite the ℓ_0 not being a norm (as it does not satisfy the homogeneity property), we will use this jargon throughout this paper for the sake of simplicity.

known as convolutional sparse coding [20]. Recently, several works have addressed the problem of using and training such a model in the context of image inpainting, super-resolution, and general image representation [21]–[25]. These methods usually exploit an ADMM formulation [26] while operating in the Fourier domain in order to search for the sparse codes and train the dictionary involved. Several variations have been proposed for solving the pursuit problem, yet there has been no theoretical analysis of their success.

Assume a signal is created by multiplying a sparse vector by a convolutional dictionary. In this work, we consider the following set of questions:

1) Can we guarantee the uniqueness of such a global (convolutional) sparse vector?
2) Can global pursuit algorithms, such as the ones suggested in recent works, be guaranteed to find the true underlying sparse code, and if so, under which conditions?
3) Can we guarantee a stability of the sparse approximation problem, and a stability of corresponding pursuit methods in a noisy regime?; And
4) Can we solve the global pursuit by restricting the process to local pursuit operations?

A naïve approach to address such theoretical questions is to apply the fairly extensive results for sparse representation and compressed sensing to the above defined model [27]. However, as we will show throughout this paper, this strategy provides nearly useless results and bounds from a global perspective. Therefore, there exists a true need for a deeper and alternative analysis of the sparse coding problem in the convolutional case which would yield meaningful bounds.

In this work, we will demonstrate the futility of the ℓ_0-norm in providing meaningful bounds in the convolutional model. This, in turn, motivates us to propose a new localized measure – the $\ell_{0,\infty}$ norm. Based on it, we redefine our pursuit into a problem that operates locally while thinking globally. To analyze this problem, we extend useful concepts, such as the Spark and mutual coherence, to the convolutional setting. We then provide claims for uniqueness of solutions and for the success of pursuit methods in the noiseless case, both for greedy algorithms and convex relaxations. Based on these theoretical foundations, we then extend our analysis to a more practical scenario, handling noisy data and model deviations. We generalize and tie past theoretical constructions, such as the Restricted Isometry Property (RIP) [28] and the Exact Recovery Condition (ERC) [29], to the convolutional framework proving the stability of this model in this case as well.

This paper is organized as follows. We begin by reviewing the unconstrained global (traditional) sparse representation model in Section II, followed by a detailed description of the convolutional structure in Section III. Section IV briefly motivates the need of a thorough analysis of this model, which is then provided in Section V. We introduce additional mathematical tools in Section VI, which provide further insight into the convolutional model. The noisy scenario is then considered in Section VII and analyzed in Section VIII, where we assume the global signal to be contaminated with norm-bounded error. We then bridge the local and global models in Section IX, proposing local algorithmic solutions to tackle the convolutional pursuit. Finally, we conclude this work in Section X, proposing exciting future directions.

II. The Global Sparse Model – Preliminaries

Consider the constrained P_0 problem, a special case of Eq. (1), given by

$$(P_0): \quad \min_{\mathbf{\Gamma}} \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \mathbf{D}\mathbf{\Gamma} = \mathbf{X}. \tag{2}$$

Several results have shed light on the theoretical aspects of this problem, claiming a unique solution under certain circumstances. These guarantees are given in terms of properties of the dictionary $\mathbf{D}$, such as the *Spark*, defined as the minimum number of linearly dependent columns (atoms) in $\mathbf{D}$ [30]. Formally,

$$\sigma(\mathbf{D}) = \min_{\mathbf{\Gamma}} \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \mathbf{D}\mathbf{\Gamma} = \mathbf{0},\ \mathbf{\Gamma} \neq \mathbf{0}.$$

Based on this property, a solution obeying $\|\mathbf{\Gamma}\|_0 < \sigma(\mathbf{D})/2$ is necessarily the sparsest one [30]. Unfortunately, this bound is of little practical use, as computing the Spark of a matrix is a combinatorial problem – and infeasible in practice.

Another guarantee is given in terms of the *mutual coherence* of the dictionary, $\mu(\mathbf{D})$, which quantifies the similarity of atoms in the dictionary. Assuming hereafter that $\|\mathbf{d}_i\|_2 = 1\,\forall i$, this was defined in [30] as:

$$\mu(\mathbf{D}) = \max_{i \neq j} |\mathbf{d}_i^T \mathbf{d}_j|.$$

A relation between the Spark and the mutual coherence was also shown in [30], stating that $\sigma(\mathbf{D}) \geq 1 + \frac{1}{\mu(\mathbf{D})}$. This, in turn, enabled the formulation of a practical uniqueness bound guaranteeing that $\mathbf{\Gamma}$ is the unique solution of the P_0 problem if $\|\mathbf{\Gamma}\|_0 < \frac{1}{2}(1 + 1/\mu(\mathbf{D}))$.

Solving the P_0 problem is NP-hard in general. Nevertheless, its solution can be approximated by either greedy pursuit algorithms, such as the Orthogonal Matching Pursuit (OMP) [31], [32], or convex relaxation approaches like Basis Pursuit (BP) [33]. Despite the difficulty of this problem, these methods (and other similar ones) have been proven to recover the true solution if $\|\mathbf{\Gamma}\|_0 < \frac{1}{2}(1 + 1/\mu(\mathbf{D}))$ [29], [30], [34], [35].

When dealing with natural signals, the P_0 problem is often relaxed to consider model deviations as well as measurement noise. In this set-up one assumes $\mathbf{Y} = \mathbf{D}\mathbf{\Gamma} + \mathbf{E}$, where $\mathbf{E}$ is a nuisance vector of bounded energy, $\|\mathbf{E}\|_2 \leq \epsilon$. The corresponding recovery problem can then be stated as follows:

$$(P_0^\epsilon): \quad \min_{\mathbf{\Gamma}} \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \|\mathbf{D}\mathbf{\Gamma} - \mathbf{Y}\|_2 \leq \epsilon.$$

Unlike the noiseless case, given a solution to the above problem, one can not claim its uniqueness in solving the P_0^ϵ problem but instead can guarantee that it will be close enough to the true vector $\mathbf{\Gamma}$ that generated the signal $\mathbf{Y}$. This kind of stability results have been derived in recent years by leveraging the Restricted Isometry Property (RIP) [28]. A matrix $\mathbf{D}$ is said to have a k-RIP with constant δ_k if this is the smallest quantity such that

$$(1 - \delta_k)\|\mathbf{\Gamma}\|_2^2 \leq \|\mathbf{D}\mathbf{\Gamma}\|_2^2 \leq (1 + \delta_k)\|\mathbf{\Gamma}\|_2^2,$$

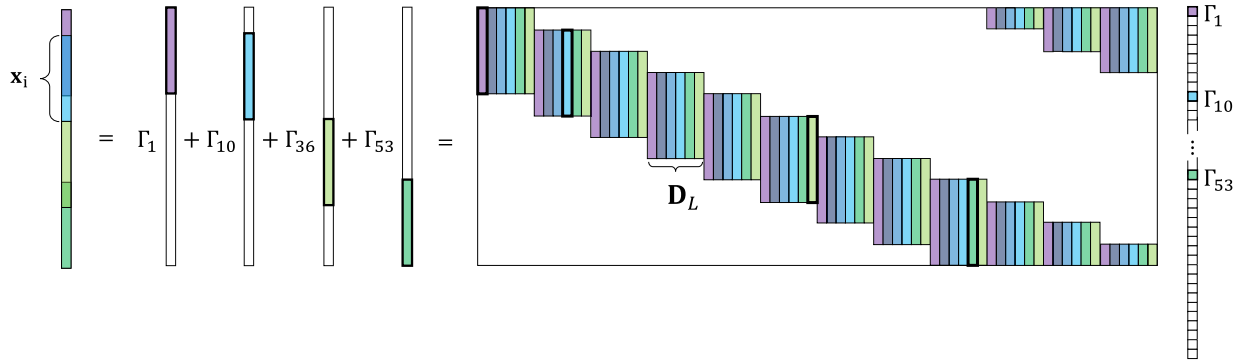


Fig. 1. The convolutional model description, and its composition in terms of the local dictionary $\mathbf{D}_L$.

for every $\mathbf{\Gamma}$ satisfying $\|\mathbf{\Gamma}\|_0 = k$. Based on this property, it was shown that assuming $\mathbf{\Gamma}$ is sparse enough, the distance between $\mathbf{\Gamma}$ and all other solutions to the P_0^ϵ problem is bounded [27]. Similar stability claims can be formulated in terms of the mutual coherence also, by exploiting its relationship with the RIP property [27].

Success guarantees of practical algorithms, such as the Orthogonal Matching Pursuit (OMP) and the Basis Pursuit Denoising (BPDN), have also been derived under this regime. In the same spirit of the aforementioned stability results, the work in [34] showed that these methods recover a solution close to the true sparse vector as long as some sparsity constraint, relying on the mutual coherence of the dictionary, is met.

Another useful property for analyzing the success of pursuit methods, initially proposed in [29], is the Exact Recovery Condition (ERC). Formally, one says that the ERC is met for a support $\mathcal{T}$ with a constant θ whenever

$$\theta = 1 - \max_{i \notin \mathcal{T}} \|\mathbf{D}_{\mathcal{T}}^{\dagger} \mathbf{d}_i\|_1 > 0,$$

where we have denoted by $\mathbf{D}_{\mathcal{T}}^{\dagger}$ the Moore-Penrose pseudoinverse of the dictionary restricted to support $\mathcal{T}$, and $\mathbf{d}_i$ refers to the ith atom in $\mathbf{D}$. Assuming the above is satisfied, the stability of both the OMP and BP was proven in [36]. Moreover, in an effort to provide a more intuitive result, the ERC was shown to hold whenever the total number of non-zeros in $\mathcal{T}$ is less than a certain number, which is a function of the mutual coherence.

III. The Convolutional Sparse Model

Consider now the global dictionary to be a concatenation of m banded Circulant matrices,[2] where each such matrix has a band of width $n \ll N$. As such, by simple permutation of its columns, such a dictionary consists of all shifted versions of a *local* dictionary $\mathbf{D}_L$ of size $n \times m$. This model is commonly known as Convolutional Sparse Representation [20], [22], [37]. Hereafter, whenever we refer to the global dictionary $\mathbf{D}$, we assume it has this structure. Assume a signal $\mathbf{X}$ to be generated as $\mathbf{D\Gamma}$. In Fig. 1 we describe such a global signal, its corresponding dictionary that is of size $N \times mN$ and its sparse representation, of length mN. We note that $\mathbf{\Gamma}$ is built of N distinct and independent sparse parts, each of length m, which we will refer to as the local sparse vectors $\boldsymbol{\alpha}_i$.

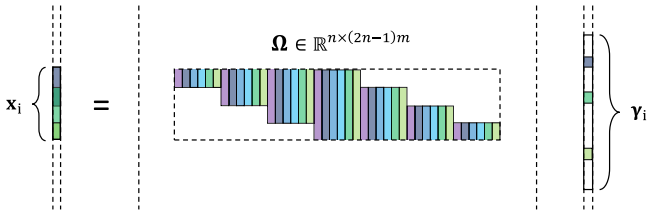


Fig. 2. Stripe Dictionary.

Consider a sub-system of equations extracted from $\mathbf{X} = \mathbf{D\Gamma}$ by multiplying this system by the patch extraction[3] operator $\mathbf{R}_i \in \mathbb{R}^{n \times N}$. The resulting system is $\mathbf{x}_i = \mathbf{R}_i\mathbf{X} = \mathbf{R}_i\mathbf{D\Gamma}$, where $\mathbf{x}_i$ is a patch of length n extracted from $\mathbf{X}$ from location i. Observe that in the set of extracted rows, $\mathbf{R}_i\mathbf{D}$, there are only $(2n-1)m$ columns that are non-trivially zero. Define the operator $\mathbf{S}_i \in \mathbb{R}^{(2n-1)m \times mN}$ as a columns' selection operator,[4] such that $\mathbf{R}_i\mathbf{D}\mathbf{S}_i^T$ preserves all the non-zero columns in $\mathbf{R}_i\mathbf{D}$. Thus, the subset of equations we get is essentially

$$\mathbf{x}_i = \mathbf{R}_i\mathbf{X} = \mathbf{R}_i\mathbf{D\Gamma} = \mathbf{R}_i\mathbf{D}\mathbf{S}_i^T\mathbf{S}_i\mathbf{\Gamma}. \tag{3}$$

Definition 1: Given a global sparse vector $\mathbf{\Gamma}$, define $\boldsymbol{\gamma}_i = \mathbf{S}_i\mathbf{\Gamma}$ as its ith stripe representation.

Note that a stripe $\boldsymbol{\gamma}_i$ can be also seen as a group of $2n-1$ adjacent local sparse vectors $\boldsymbol{\alpha}_j$ of length m from $\mathbf{\Gamma}$, centered at location $\boldsymbol{\alpha}_i$.

Definition 2: Consider a convolutional dictionary $\mathbf{D}$ defined by a local dictionary $\mathbf{D}_L$ of size $n \times m$. Define the stripe dictionary $\mathbf{\Omega}$ of size $n \times (2n-1)m$, as the one obtained by extracting n consecutive rows from $\mathbf{D}$, followed by the removal of its zero columns, namely $\mathbf{\Omega} = \mathbf{R}_i\mathbf{D}\mathbf{S}_i^T$.

Observe that $\mathbf{\Omega}$, depicted in Fig. 2, is independent of i, being the same for all locations due to the union-of-Circulant-matrices structure of $\mathbf{D}$. In other words, the shift invariant property is sat-

[2]Each of these matrices is constructed by shifting a single column, supported on n subsequent entries, to all possible shifts. This choice of Circulant matrices comes to alleviate boundary problems.

[3]Denoting by $\mathbf{0}_{(a\times b)}$ a zeros matrix of size $a \times b$, and $\mathbf{I}_{(n\times n)}$ an identity matrix of size $n \times n$, then $\mathbf{R}_i = [\mathbf{0}_{(n\times(i-1))}, \mathbf{I}_{(n\times n)}, \mathbf{0}_{(n\times(N-i-n+1))}]$.

[4]An analogous definition can be written for this operator as well.

isfied for this model – all patches share the same stripe dictionary in their construction. Armed with the above two definitions, Eq. (3) simply reads $\mathbf{x}_i = \mathbf{\Omega}\boldsymbol{\gamma}_i$.

From a different perspective, one can synthesize the signal $\mathbf{X}$ by considering $\mathbf{D}$ as a concatenation of N vertical stripes of size $N \times m$ (see Fig. 1), where each can be represented as $\mathbf{R}_i^T \mathbf{D}_L$. In other words, the vertical stripe is constructed by taking the small and local dictionary $\mathbf{D}_L$ and positioning it in the ith row. The same partitioning applies to $\mathbf{\Gamma}$, leading to the $\boldsymbol{\alpha}_i$ ingredients. Thus,

$$\mathbf{X} = \sum_i \mathbf{R}_i^T \mathbf{D}_L \boldsymbol{\alpha}_i.$$

Since $\boldsymbol{\alpha}_i$ play the role of local sparse vectors, $\mathbf{D}_L\boldsymbol{\alpha}_i$ are reconstructed patches (which are not the same as $\mathbf{x}_i = \mathbf{\Omega}\boldsymbol{\gamma}_i$), and the sum above proposes a patch averaging approach as practiced in several works [8], [14], [19]. This formulation provides another local interpretation of the convolutional model.

Yet a third interpretation of the very same signal construction can be suggested, in which the signal is seen as resulting from a sum of local/small atoms which appear in a small number of locations throughout the signal. This can be formally expressed as

$$\mathbf{X} = \sum_{i=1}^{m} \mathbf{d}_i * \mathbf{z}_i,$$

where the vectors $\mathbf{z}_i \in \mathbb{R}^N$ are sparse maps encoding the location and coefficients convolved with the ith atom [20]. In our context, $\mathbf{\Gamma}$ is simply the interlaced concatenation of all $\mathbf{z}_i$.

This model (adopting the last convolutional interpretation) has received growing attention in recent years in various applications. In [38] a convolutional sparse coding framework was used for pattern detection in images and the analysis of instruments in music signals, while in [39] it was used for the reconstruction of 3D trajectories. The problem of learning the local dictionary $\mathbf{D}_L$ was also studied in several works [24], [37], [40]–[42]. Different methods have been proposed for solving the convolutional sparse coding problem under an ℓ_1-norm penalty. Commonly, these methods rely on the ADMM algorithm [26], exploiting multiplications of vectors by the global dictionary in the Fourier domain in order to reduce the computational cost involved [37]. An alternative is the deployment of greedy algorithms of the Matching Pursuit family [5], which suggest an ℓ_0 constraint on the global sparse vector. The reader is referred to the work of [24] and references therein for a thorough discussion on these methods. In essence, all the above works are solutions to the minimization of a global pursuit under the convolutional structure. As a result, the theoretical results in our work will apply to these methods, providing guarantees for the recovery of the underlying sparse vectors.

IV. From Global to Local Analysis

Consider a sparse vector $\mathbf{\Gamma}$ of size mN which represents a global (convolutional) signal. Assume further that this vector has a few $k \ll N$ non-zeros. If these were to be clustered together in a given stripe $\boldsymbol{\gamma}_i$, the local patch corresponding to this stripe would be very complex, and pursuit methods would likely fail in recovering it. On the contrary, if these k non-zeros are spread all throughout the vector $\mathbf{\Gamma}$, this would clearly imply much simpler local patches, facilitating their successful recovery. This simple example comes to show the futility of the traditional global ℓ_0-norm in assessing the success of convolutional pursuits, and it will be the pillar of our intuition throughout our work.

A. The $\ell_{0,\infty}$ Norm and the $P_{0,\infty}$ Problem

Let us now introduce a measure that will provide a local notion of sparsity within a global sparse vector.

Definition 3: Define the $\ell_{0,\infty}$ pseudo-norm of a global sparse vector $\mathbf{\Gamma}$ as

$$\|\mathbf{\Gamma}\|_{0,\infty} = \max_i \|\boldsymbol{\gamma}_i\|_0.$$

In words, this quantifies the number of non-zeros in the densest stripe $\boldsymbol{\gamma}_i$ of the global $\mathbf{\Gamma}$. This is equivalent to extracting all stripes from the global sparse vector $\mathbf{\Gamma}$, arranging them column-wise into a matrix $\mathbf{A}$ and applying the usual $\|\mathbf{A}\|_{0,\infty}$ norm – thus, the name. By constraining the $\ell_{0,\infty}$ norm to be low, we are essentially limiting all stripes $\boldsymbol{\gamma}_i$ to be sparse, and their corresponding patches $\mathbf{R}_i\mathbf{X}$ to have a sparse representation under a shift-invariant local dictionary $\mathbf{\Omega}$. This is one of the underlying assumptions in many signal and image processing algorithms. As for properties of this norm, similar to ℓ_0 case, in the $\ell_{0,\infty}$ the non-negativity and triangle inequality properties hold, while homogeneity does not.

Armed with the above definition, we now move to define the $P_{0,\infty}$ problem:

$$(P_{0,\infty}): \quad \min_{\mathbf{\Gamma}} \quad \|\mathbf{\Gamma}\|_{0,\infty} \text{ s.t. } \mathbf{D}\mathbf{\Gamma} = \mathbf{X}.$$

When dealing with a global signal, instead of solving the P_0 problem (defined in Eq. (2)) as is commonly done, we aim to solve the above defined objective instead. The key difference is that we are not limiting the overall number of zeros in $\mathbf{\Gamma}$, but rather putting a restriction on its local density.

B. Global versus Local Bounds

As mentioned previously, theoretical bounds are often given in terms of the mutual coherence of the dictionary. In this respect, a lower bound on this value is much desired. In the case of the convolution sparse model, this value quantifies not only the correlation between the atoms in $\mathbf{D}_L$, but also the correlation between their shifts. Though in a different context, a bound for this value was derived in [43], and it is given by

$$\mu(\mathbf{D}) \geq \sqrt{\frac{m-1}{m(2n-1)-1}}. \tag{4}$$

For a large value of m, one obtains that the best possible coherence is $\mu(\mathbf{D}) \approx \frac{1}{\sqrt{2n}}$. This implies that if we are to apply BP or OMP to recover the sparsest $\mathbf{\Gamma}$ that represents $\mathbf{X}$, the classical sparse approximation results [1] would allow merely $O(\sqrt{n})$ non-zeros in **all** $\mathbf{\Gamma}$, for any N, no matter how long $\mathbf{X}$ is! As we shall see next, the situation is not as grave as it may seem, due to

TABLE I
SUMMARY OF NOTATIONS USED THROUGHOUT THE PAPER

N	:	length of the global signal.
n	:	size of a local atom or a local signal patch.
m	:	number of unique local atoms (filters) or the number of Circulant matrices.
$\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{E}$	:	global signals of length N, where generally $\mathbf{Y} = \mathbf{X} + \mathbf{E}$.
$\mathbf{D}$	:	global dictionary of size $N \times mN$.
$\mathbf{\Gamma}$ and $\mathbf{\Delta}$	:	global sparse vectors of length mN.
Γ_i and Δ_i	:	the i^{th} entry in $\mathbf{\Gamma}$ and $\mathbf{\Delta}$, respectively.
$\mathbf{D}_L$	:	local dictionary of size $n \times m$.
$\mathbf{\Omega}$	:	stripe dictionary of size $n \times (2n-1)m$, which contains all possible shifts of $\mathbf{D}_L$.
$\boldsymbol{\alpha}_i$	:	local sparse code of size m.
$\boldsymbol{\gamma}_i$ and $\boldsymbol{\delta}_i$	:	a stripe of length $(2n-1)m$ extracted from the global vectors $\mathbf{\Gamma}$ and $\mathbf{\Delta}$, respectively.
$\boldsymbol{\gamma}_{i,s}$ and $\boldsymbol{\delta}_{i,s}$	:	a local sparse vector of length m which corresponds to the s^{th} portion inside $\boldsymbol{\gamma}_i$ and $\boldsymbol{\delta}_i$, respectively.

our migration from P_0 to $P_{0,\infty}$. Leveraging the previous definitions, we will provide global recovery guarantees that will have a local flavor, and the bounds will be given in terms of the number of non-zeros in the densest stripe. This way, we will show that the guarantee conditions can be significantly enhanced to $O(\sqrt{n})$ non-zeros *locally* rather than *globally*.

V. THEORETICAL STUDY OF IDEAL SIGNALS

As motivated in the previous section, the concerns of uniqueness, recovery guarantees and stability of sparse solutions in the convolutional case require special attention. We now formally address these questions by following the path taken in [27], carefully generalizing each and every statement to the global-local model discussed here.

Before proceeding onto theoretical grounds, we briefly summarize, for the convenience of the reader, all notations used throughout this work in Table I. Note the somewhat unorthodox choice of capital letters for global vectors and lowercase for local ones.

A. Uniqueness and Stripe-Spark

Just as it was initially done in the general sparse model, one might ponder about the uniqueness of the sparsest representation in terms of the $\ell_{0,\infty}$ norm. More precisely, does a unique solution to the $P_{0,\infty}$ problem exist? and under which circumstances? In order to answer these questions we shall first extend our mathematical tools, in particular the characterization of the dictionary, to the convolutional scenario.

In Section II we recalled the definition of the Spark of a general dictionary $\mathbf{D}$. In the same spirit, we propose the following:

Definition 4: Define the Stripe-Spark of a convolutional dictionary $\mathbf{D}$ as

$$\sigma_\infty(\mathbf{D}) = \min_{\mathbf{\Delta}} \quad \|\mathbf{\Delta}\|_{0,\infty} \ \text{ s.t. } \ \mathbf{\Delta} \neq 0,\ \mathbf{D}\mathbf{\Delta} = 0.$$

In words, the Stripe-Spark is defined by the sparsest non-zero vector, in terms of the $\ell_{0,\infty}$ norm, in the null space of $\mathbf{D}$. Next, we use this definition in order to formulate an uncertainty and a uniqueness principle for the $P_{0,\infty}$ problem that emerges from it. The proof of this and the following theorems are described in detail in the Supplementary Material.

Theorem 5 (Uncertainty and uniqueness using Stripe-Spark): Let $\mathbf{D}$ be a convolutional dictionary. If a solution $\mathbf{\Gamma}$ obeys $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\sigma_\infty$, then this is necessarily the global optimum for the $P_{0,\infty}$ problem for the signal $\mathbf{D\Gamma}$.

B. Lower Bounding the Stripe-Spark

In general, and similar to the Spark, calculating the Stripe-Spark is computationally intractable. Nevertheless, one can bound its value using the global mutual coherence defined in Section II. Before presenting such bound, we formulate and prove a Lemma that will aid our analysis throughout this paper.

Lemma 1: Consider a convolutional dictionary $\mathbf{D}$, with mutual coherence $\mu(\mathbf{D})$, and a support $\mathcal{T}$ with $\ell_{0,\infty}$ norm[5] equal to k. Let $\mathbf{G}^{\mathcal{T}} = \mathbf{D}_{\mathcal{T}}^T\mathbf{D}_{\mathcal{T}}$, where $\mathbf{D}_{\mathcal{T}}$ is the matrix $\mathbf{D}$ restricted to the columns indicated by the support $\mathcal{T}$. Then, the eigenvalues of this Gram matrix, given by $\lambda_i(\mathbf{G}^{\mathcal{T}})$, are bounded by

$$1 - (k-1)\mu(\mathbf{D}) \ \le \ \lambda_i\left(\mathbf{G}^{\mathcal{T}}\right) \ \le \ 1 + (k-1)\mu(\mathbf{D}).$$

Proof: From Gerschgorin's theorem, the eigenvalues of the Gram matrix $\mathbf{G}^{\mathcal{T}}$ reside in the union of its Gerschgorin circles. The jth circle, corresponding to the jth row of $\mathbf{G}^{\mathcal{T}}$, is centered at the point $\mathbf{G}^{\mathcal{T}}_{j,j}$ (belonging to the Gram's diagonal) and its radius equals the sum of the absolute values of the off-diagonal entries; i.e., $\sum_{i,i\neq j}|\mathbf{G}^{\mathcal{T}}_{j,i}|$. Notice that both indices i, j correspond to atoms in the support $\mathcal{T}$. Because the atoms are normalized, $\forall\ j,\ \mathbf{G}^{\mathcal{T}}_{j,j} = 1$, implying that all Gershgorin disks are centered at 1. Therefore, all eigenvalues reside inside the circle with the largest radius. Formally,

$$\left|\lambda_i\left(\mathbf{G}^{\mathcal{T}}\right) - 1\right| \le \max_j \sum_{i,i\neq j}\left|\mathbf{G}^{\mathcal{T}}_{j,i}\right| = \max_j \sum_{\substack{i,i\neq j\\ i,j\in\mathcal{T}}}\left|\mathbf{d}_j^T\mathbf{d}_i\right|. \quad (5)$$

On the one hand, from the definition of the mutual coherence, the inner product between atoms that are close enough to overlap is bounded by $\mu(\mathbf{D})$. On the other hand, the product $\mathbf{d}_j^T\mathbf{d}_i$ is zero for atoms $\mathbf{d}_i$ too far from $\mathbf{d}_j$ (i.e., out of the stripe centered at the jth atom). Therefore, we obtain:

$$\sum_{\substack{i,i\neq j\\ i,j\in\mathcal{T}}}\left|\mathbf{d}_j^T\mathbf{d}_i\right| \le (k-1)\ \mu(\mathbf{D}),$$

where k is the maximal number of non-zero elements in a stripe, defined previously as the $\ell_{0,\infty}$ norm of $\mathcal{T}$. Note that we have subtracted 1 from k because we must omit the entry on the diagonal. Putting this back in Eq. (5), we obtain

$$\left|\lambda_i\left(\mathbf{G}^{\mathcal{T}}\right) - 1\right| \le \max_j \sum_{\substack{i,i\neq j\\ i,j\in\mathcal{T}}}\left|\mathbf{d}_j^T\mathbf{d}_i\right| \le \ (k-1)\ \mu(\mathbf{D}).$$

From this we obtain the desired claim. ■

We now dive into the next theorem, whose proof relies on the above Lemma.

[5]Note that specifying the $\ell_{0,\infty}$ of a support rather than a sparse vector is a slight abuse of notation, that we will nevertheless use for the sake of simplicity.

Theorem 6 (Lower bounding the Stripe-Spark via the mutual coherence): For a convolutional dictionary $\mathbf{D}$ with mutual coherence $\mu(\mathbf{D})$, the Stripe-Spark can be lower-bounded by

$$\sigma_\infty(\mathbf{D}) \geq 1 + \frac{1}{\mu(\mathbf{D})}.$$

Using the above derived bound and the uniqueness based on the Stripe-Spark we can now formulate the following theorem:

Theorem 7 (Uniqueness using mutual coherence): Let $\mathbf{D}$ be a convolutional dictionary with mutual coherence $\mu(\mathbf{D})$. If a solution $\mathbf{\Gamma}$ obeys $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D})})$, then this is necessarily the sparsest (in terms of $\ell_{0,\infty}$ norm) solution to $P_{0,\infty}$ with the signal $\mathbf{D\Gamma}$.

The proof of this claim is rather trivial, noting that if $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D})})$, then necessarilly $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\sigma_\infty$, and so from Theorem 5 $\mathbf{\Gamma}$ is unique.

At the end of Section IV we mentioned that for $m \gg 1$, the classical analysis would allow an order of $O(\sqrt{n})$ non-zeros all over the vector $\mathbf{\Gamma}$, regardless of the length of the signal N. In light of the above theorem, in the convolutional case, the very same quantity of non-zeros is allowed locally per stripe, implying that the overall number of non-zeros in $\mathbf{\Gamma}$ grows linearly with the global dimension N.

C. Recovery Guarantees for Pursuit Methods

In this subsection, we attempt to solve the $P_{0,\infty}$ problem by employing two common, but very different, pursuit methods: the Orthogonal Matching Pursuit (OMP) and the Basis Pursuit (BP) – the reader is referred to [27] for a detailed description of these formulations and respective algorithms. Leaving aside the computational burdens of running such algorithms, which will be addressed in the second part of this work, we now consider the theoretical aspects of their success.

Previous works [29], [30] have shown that both OMP and BP succeed in finding the sparsest solution to the P_0 problem if the cardinality of the representation is known a priori to be lower than $\frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D})})$. That is, we are guaranteed to recover the underlying solution as long as the *global sparsity* is less than a certain threshold. In light of the discussion in Section IV-B, these values are pessimistic in the convolutional setting. By migrating from P_0 to the $P_{0,\infty}$ problem, we show next that both algorithms are in fact capable of recovering the underlying solutions under far weaker assumptions.

Theorem 8 (Global OMP recovery guarantee using $\ell_{0,\infty}$ norm): Given the system of linear equations $\mathbf{X} = \mathbf{D\Gamma}$, if a solution $\mathbf{\Gamma}$ exists satisfying

$$\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right),$$

then OMP is guaranteed to recover it.

Note that if we assume $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D})})$, according to our uniqueness theorem, the solution obtained by the OMP is the unique solution to the $P_{0,\infty}$ problem. Interestingly, under the same conditions the BP algorithm is guaranteed to succeed as well.

Theorem 9 (Global Basis Pursuit recovery guarantee using the $\ell_{0,\infty}$ norm): For the system of linear equations $\mathbf{D\Gamma} = \mathbf{X}$, if a solution $\mathbf{\Gamma}$ exists obeying

$$\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right),$$

then Basis Pursuit is guaranteed to recover it.

The recovery guarantees for both pursuit methods have now become *independent of the global signal dimension and sparsity*. Instead, the condition for success is given in terms of the *local* concentration of non-zeros of the global sparse vector. Moreover, the number of non-zeros allowed per stripe under the current bounds is in fact the same number previously allowed globally. As a remark, note that we have used these two algorithms in their natural form, being oblivious to the $\ell_{0,\infty}$ objective they are serving. Further work is required to develop OMP and BP versions that are aware of this specific goal, potentially benefiting from it.

D. Experiments

In this subsection we intend to provide numerical results that corroborate the above presented theoretical bounds. While doing so, we will shed light on the performance of the OMP and BP algorithms in practice, as compared to our previous analysis.

In [44] an algorithm was proposed to construct a local dictionary such that all its aperiodic auto-correlations and cross-correlations are low. This, in our context, means that the algorithm attempts to minimize the mutual coherence of the dictionary $\mathbf{D}_L$ and all of its shifts, decreasing the global mutual coherence as a result. We use this algorithm to numerically build a dictionary consisting of two atoms ($m = 2$) with patch size $n = 64$. The theoretical lower bound on the $\mu(\mathbf{D})$ presented in Eq. (4) under this setting is approximately 0.063, and we manage to obtain a mutual coherence of 0.09 using the aforementioned method. With these atoms we construct a convolutional dictionary with global atoms of length $N = 640$.

Once the dictionary is fixed, we generate sparse vectors with random supports of (global) cardinalities in the range $[1, 300]$. The non-zero entries are drawn from random independent and identically-distributed Gaussians with mean equal to zero and variance equal to one. Given these sparse vectors, we compute their corresponding global signals and attempt to recover them using the global OMP and BP. We perform 500 experiments per each cardinality and present the probability of success as a function of the representation's $\ell_{0,\infty}$ norm. We define the success of the algorithm as the full recovery of the true sparse vector. The results for the experiment are presented in Fig. 3. The theorems provided in the previous subsection guarantee the success of both OMP and BP as long as the $\|\mathbf{\Gamma}\|_{0,\infty} \leq 6$.

As can be seen from these results, the theoretical bound is far from being tight. However, in the traditional sparse representation model the corresponding bounds have the same loose flavor [1]. This kind of results is in fact expected when using such a worst-case analysis. Tighter bounds could likely be obtained by a probabilistic study, which we leave for future work.

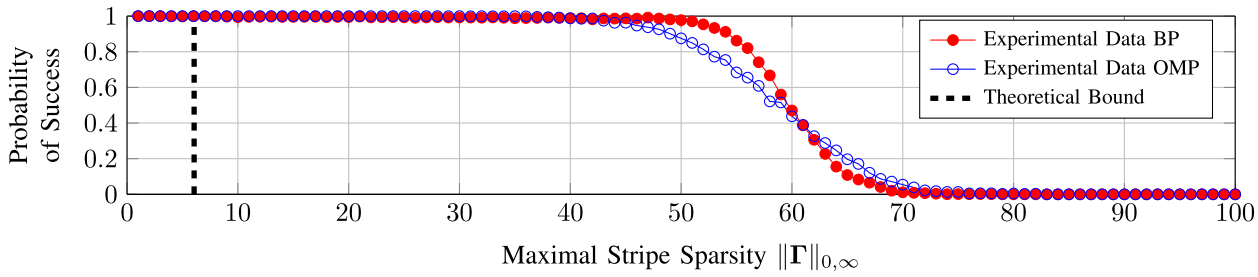


Fig. 3. Probability of success of OMP and BP at recovering the true convolutional sparse code. The theoretical guarantee is presented on the same graph.

VI. Shifted Mutual Coherence and Stripe Coherence

When considering the mutual coherence $\mu(\mathbf{D})$, one needs to look at the maximal correlation between every pair of atoms in the global dictionary. One should note, however, that atoms having a non-zero correlation must have overlapping supports, and $\mu(\mathbf{D})$ provides a bound for these values independently of the amount of overlap. One could go beyond this characterization of the convolutional dictionary by a single value and propose to bound all the inner products between atoms for a *given shift*. As a motivation, in several applications one can assume that signals are built from local atoms separated by some minimal lag, or shift. In radio communications, for example, such a situation appears when there exists a minimal time between consecutive transmissions on the same channel [45]. In such cases, knowing how the correlation between the atoms depends on their shifts is fundamental for the design of the dictionary, its utilization and its theoretical analysis.

In this section we briefly explore this direction of analysis, introducing a stronger characterization of the convolutional dictionary, termed shifted mutual coherence. By being a considerably more informative measure than the standard mutual coherence, this will naturally lead to stronger bounds. We will only present the main points of these results here for the sake of brevity; the interested reader can find a more detailed discussion on this matter in the Supplementary Material.

Recall that $\mathbf{\Omega}$ is defined as a stripe extracted from the global dictionary $\mathbf{D}$. Consider the sub-system given by $\mathbf{x}_i = \mathbf{\Omega}\boldsymbol{\gamma}_i$, corresponding to the ith patch in $\mathbf{X}$. Note that $\mathbf{\Omega}$ can be split into a set of $2n-1$ blocks of size $n \times m$, where each block is denoted by $\mathbf{\Omega}_s$, i.e.,

$$\mathbf{\Omega} = [\mathbf{\Omega}_{-n+1}, \ldots, \mathbf{\Omega}_{-1}, \mathbf{\Omega}_0, \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_{n-1}],$$

as shown previously in Fig. 2.

Definition 10: Define the shifted mutual coherence μ_s by

$$\mu_s = \max_{i,j} \quad |\langle \mathbf{d}_i^0, \mathbf{d}_j^s \rangle|,$$

where $\mathbf{d}_i^0$ is a column extracted from $\mathbf{\Omega}_0$, $\mathbf{d}_j^s$ is extracted from $\mathbf{\Omega}_s$, and we require[6] that $i \neq j$ if $s = 0$.

The above definition can be seen as a generalization of the mutual coherence for the shift-invariant local model presented in Section III. Indeed, μ_s characterizes $\mathbf{\Omega}$ just as $\mu(\mathbf{D})$ characterizes the coherence of a general dictionary. Note that if $s = 0$ the above definition boils down to the traditional mutual coherence of $\mathbf{D}_L$, i.e., $\mu_0 = \mu(\mathbf{D}_L)$. It is important to stress that the atoms used in the above definition *are normalized globally* according to $\mathbf{D}$ and not $\mathbf{\Omega}$. In the Supplementary Material we comment on several interesting properties of this measure.

Similar to $\mathbf{\Omega}$, $\boldsymbol{\gamma}_i$ can be split into a set of $2n-1$ vectors of length m, each denoted by $\boldsymbol{\gamma}_{i,s}$ and corresponding to $\mathbf{\Omega}_s$. In other words, $\boldsymbol{\gamma}_i = [\boldsymbol{\gamma}_{i,-n+1}^T, \ldots, \boldsymbol{\gamma}_{i,-1}^T, \boldsymbol{\gamma}_{i,0}^T, \boldsymbol{\gamma}_{i,1}^T, \ldots, \boldsymbol{\gamma}_{i,n-1}^T]^T$. Note that previously we denoted local sparse vectors of length m by $\boldsymbol{\alpha}_j$. Yet, we will also denote them by $\boldsymbol{\gamma}_{i,s}$ in order to emphasize the fact that they correspond to the sth shift within $\boldsymbol{\gamma}_i$. Denote the number of non-zeros in $\boldsymbol{\gamma}_i$ as n_i. We can also write $n_i = \sum_{s=-n+1}^{n-1} n_{i,s}$, where $n_{i,s}$ is the number of non-zeros in each $\boldsymbol{\gamma}_{i,s}$. With these definitions, we can now propose the following measure.

Definition 11: Define the stripe coherence as

$$\zeta(\boldsymbol{\gamma}_i) = \sum_{s=-n+1}^{n-1} n_{i,s}\, \mu_s.$$

According to this definition, each stripe has a coherence given by the sum of its non-zeros weighted by the shifted mutual coherence. As a particular case, if all k non-zeros correspond to atoms in the center sub-dictionary, $\mathbf{D}_L$, this becomes $\mu_0 k$. Note that unlike the traditional mutual coherence, this new measure depends on the location of the non-zeros in $\mathbf{\Gamma}$ – it is a function of the support of the sparse vector, and not just of the dictionary. As such, it characterizes the correlation between the atoms participating in a given stripe. In what follows, we will use the notation ζ_i for $\zeta(\boldsymbol{\gamma}_i)$.

Having formalized these tighter constructions, we now leverage them to improve the previous results. Although these theorems are generally sharper, they are harder to grasp. We begin with a recovery guarantee for the OMP and BP algorithms, followed by a discussion on their implications.

Theorem 12 (Global OMP recovery guarantee using the stripe coherence): Given the system of linear equations $\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$, if a solution $\mathbf{\Gamma}$ exists satisfying

$$\max_i \zeta_i = \max_i \sum_{s=-n+1}^{n-1} n_{i,s}\mu_s < \frac{1}{2}(1+\mu_0), \qquad (6)$$

then OMP is guaranteed to recover it.

Theorem 13 (Global BP recovery guarantee using the stripe coherence): Given the system of linear equations $\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$, if a

[6]The condition $i \neq j$ if $s = 0$ is necessary so as to avoid the inner product of an atom by itself.

solution $\mathbf{\Gamma}$ exists satisfying

$$\max_i \zeta_i = \max_i \sum_{s=-n+1}^{n-1} n_{i,s}\mu_s < \frac{1}{2}(1+\mu_0),$$

then Basis Pursuit is guaranteed to recover it.

The corresponding proofs are similar to their counterparts presented in the preceding section but require a more delicate analysis. We include the proof for the OMP variant in the Supplementary Material, and outline the main steps required to prove the BP version.

In order to provide an intuitive interpretation for these results, the above bounds can be tied to a concrete number of non-zeros per stripe. First, notice that requiring the maximal stripe coherence to be less than a certain threshold is equal to requiring the same for every stripe:

$$\forall i \quad \sum_{s=-n+1}^{n-1} n_{i,s}\mu_s < \frac{1}{2}(1+\mu_0).$$

Multiplying and dividing the left-hand side of the above inequality by n_i and rearranging the resulting expression, we obtain

$$\forall i \quad n_i < \frac{1}{2}\frac{1+\mu_0}{\sum_{s=-n+1}^{n-1}\frac{n_{i,s}}{n_i}\mu_s}.$$

Define $\bar{\mu}_i = \sum_{s=-n+1}^{n-1}\frac{n_{i,s}}{n_i}\mu_s$. Recall that $\sum_{s=-n+1}^{n-1}\frac{n_{i,s}}{n_i} = 1$ and as such $\bar{\mu}_i$ is simply the (weighted) average shifted mutual coherence in the ith stripe. Putting this definition into the above condition, the inequality becomes

$$\forall i \quad n_i < \frac{1}{2}\left(\frac{1}{\bar{\mu}_i} + \frac{\mu_0}{\bar{\mu}_i}\right).$$

Thus, the condition in (6) boils down to requiring the sparsity of all stripes to be less than a certain number. Naturally, this inequality resembles the one presented in the previous section for the OMP and BP guarantees. In the Supplementary Material we prove that under the assumption that $\mu(\mathbf{D}) = \mu_0$, the shifted mutual coherence condition is at least as strong as the original one.

VII. From Global to Local Stability Analysis

One of the cardinal motivations for this work was a series of recent practical methods addressing the convolutional sparse coding problem; and in particular, the need for their theoretical foundation. However, our results are as of yet not directly applicable to these, as we have restricted our analysis to the ideal case of noiseless signals. This is the path we undertake in the following sections, exploring the question of whether the convolutional model remains stable in the presence of noise.

Assume a clean signal $\mathbf{X}$, which admits a sparse representation $\mathbf{\Gamma}$ in terms of the convolutional dictionary $\mathbf{D}$, is contaminated with noise $\mathbf{E}$ (of bounded energy, $\|\mathbf{E}\|_2 \leq \epsilon$) to create $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$. Given this noisy signal, one could propose to recover the true representation $\mathbf{\Gamma}$, or a vector close to it, by solving the P_0^ϵ problem. In this context, as mentioned in the previous section, several theoretical guarantees have been proposed in the literature. As an example, consider the stability results presented in the seminal work of [34]. Therein, it was shown that assuming the total number of non-zeros in $\mathbf{\Gamma}$ is less than $\frac{1}{2}(1+\frac{1}{\mu(\mathbf{D})})$, the distance between the solution to the P_0^ϵ problem, $\overline{\mathbf{\Gamma}}$, and the true sparse vector, $\mathbf{\Gamma}$, satisfies

$$\|\overline{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_2^2 \leq \frac{4\epsilon^2}{1-\mu(\mathbf{D})(2\|\mathbf{\Gamma}\|_0 - 1)}. \tag{7}$$

In the context of our convolutional setting, however, this result provides a weak bound as it constrains the total number of non-zeros to be below a certain threshold, which scales with the local filter size n.

We now re-define the P_0^ϵ problem into a different one, capturing the convolutional structure by relying on the $\ell_{0,\infty}$ norm instead. Consider the problem:

$$(P_{0,\infty}^\epsilon): \quad \min_{\mathbf{\Gamma}} \quad \|\mathbf{\Gamma}\|_{0,\infty} \text{ s.t. } \|\mathbf{Y} - \mathbf{D\Gamma}\|_2 \leq \epsilon.$$

In words, given a noisy measurement $\mathbf{Y}$, we seek for the $\ell_{0,\infty}$-sparsest representation vector that explains this signal up to an ϵ error. In what follows, we address the theoretical aspects of this problem and, in particular, study the stability of its solutions and practical yet secured ways for retrieving them.

VIII. Theoretical Analysis of Corrupted Signals

A. Stability of the $P_{0,\infty}^\epsilon$ Problem

As expected, one cannot guarantee the uniqueness of the solution to the $P_{0,\infty}^\epsilon$ problem, as was done for the $P_{0,\infty}$. Instead, in this subsection we shall provide a stability claim that guarantees the found solution to be close to the underlying sparse vector that generated $\mathbf{Y}$. In order to provide such an analysis, we commence by arming ourselves with the necessary mathematical tools.

Definition 14: Let $\mathbf{D}$ be a convolutional dictionary. Consider all the sub matrices $\mathbf{D}_\mathcal{T}$, obtained by restricting the dictionary $\mathbf{D}$ to a support $\mathcal{T}$ with an $\ell_{0,\infty}$ norm equal to k. Define δ_k as the smallest quantity such that

$$\forall \mathbf{\Delta} \quad (1-\delta_k)\|\mathbf{\Delta}\|_2^2 \leq \|\mathbf{D}_\mathcal{T}\mathbf{\Delta}\|_2^2 \leq (1+\delta_k)\|\mathbf{\Delta}\|_2^2$$

holds true for any choice of the support. Then, $\mathbf{D}$ is said to satisfy k-SRIP (Stripe-RIP) with constant δ_k.

Given a matrix $\mathbf{D}$, similar to the Stripe-Spark, computing the SRIP is hard or practically impossible. Thus bounding it using the mutual coherence is of practical use.

Theorem 15 (Upper bounding the SRIP via the mutual coherence): For a convolutional dictionary $\mathbf{D}$ with global mutual coherence $\mu(\mathbf{D})$, the SRIP can be upper-bounded by

$$\delta_k \leq (k-1)\mu(\mathbf{D}).$$

Assume a sparse vector $\mathbf{\Gamma}$ is multiplied by $\mathbf{D}$ and then contaminated by a vector $\mathbf{E}$, generating the signal $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$, such that $\|\mathbf{Y} - \mathbf{D\Gamma}\|_2^2 \leq \epsilon^2$. Suppose we solve the $P_{0,\infty}^\epsilon$ problem and obtain a solution $\hat{\mathbf{\Gamma}}$. How close is this solution to the original $\mathbf{\Gamma}$? The following theorem provides an answer to this question.

Theorem 16 (Stability of the solution to the $P_{0,\infty}^\epsilon$ problem): Consider a sparse vector $\mathbf{\Gamma}$ such that $\|\mathbf{\Gamma}\|_{0,\infty} = k < \frac{1}{2}(1+$

$\frac{1}{\mu(\mathbf{D})}$), and a convolutional dictionary $\mathbf{D}$ satisfying the SRIP property for $\ell_{0,\infty} = 2k$ with coefficient δ_{2k}. Then, the distance between the true sparse vector $\mathbf{\Gamma}$ and the solution to the $P^{\epsilon}_{0,\infty}$ problem $\hat{\mathbf{\Gamma}}$ is bounded by

$$\|\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\|_2^2 \leq \frac{4\epsilon^2}{1-\delta_{2k}} \leq \frac{4\epsilon^2}{1-(2k-1)\mu(\mathbf{D})}. \tag{8}$$

One should wonder if the new guarantee presents any advantage when compared to the bound based on the traditional RIP. Looking at the original stability claim for the global system, as discussed in Section IV, the reader should compare the assumptions on the sparse vector $\mathbf{\Gamma}$, as well as the obtained bounds on the distance between the estimates and the original vector. The stability claim in the P^{ϵ}_0 problem is valid under the condition

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

In contrast, the stability claim presented above holds whenever

$$\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

This allows for significantly more non-zeros in the global signal. Furthermore, as long as the above hold, and comparing Eq. (7) and (8), we have that

$$\frac{4\epsilon^2}{1-(2\|\mathbf{\Gamma}\|_{0,\infty}-1)\mu(\mathbf{D})} \ll \frac{4\epsilon^2}{1-(2\|\mathbf{\Gamma}\|_0-1)\mu(\mathbf{D})},$$

since generally $\|\mathbf{\Gamma}\|_{0,\infty} \ll \|\mathbf{\Gamma}\|_0$. This inequality implies that the above developed bound is (usually much) lower than the traditional one. In other words, the bound on the distance to the true sparse vector is much tighter and far more informative under the $\ell_{0,\infty}$ setting.

B. Stability Guarantee of OMP

Hitherto, we have shown that the solution to the $P^{\epsilon}_{0,\infty}$ problem will be close to the true sparse vector $\mathbf{\Gamma}$. However, it is also important to know whether this solution can be approximated by pursuit algorithms. In this subsection, we address such a question for the OMP, extending the analysis presented to the noisy setting.

In [34], a claim was provided for the OMP, guaranteeing the recovery of the true support of the underlying solution if

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) - \frac{1}{\mu(\mathbf{D})} \cdot \frac{\epsilon}{|\Gamma_{min}|},$$

$|\Gamma_{min}|$ being the minimal absolute value of a (non-zero) coefficients in $\mathbf{\Gamma}$. This result comes to show the importance of both the sparsity of $\mathbf{\Gamma}$ and the signal-to-noise ratio, which relates to the term $\epsilon/|\Gamma_{min}|$. In the context of our convolutional setting, this result provides a weak bound for two different reasons. First, the above bound restricts the total number of non-zeros in the representation of the signal. From Section V, it is natural to seek for an alternative condition for the success of this pursuit relying on the $\ell_{0,\infty}$ norm instead. Second, notice that the rightmost term in the above bound divides the global error energy by the minimal coefficient (in absolute value) in $\mathbf{\Gamma}$. In the convolutional scenario, the energy of the error ϵ is a *global* quantity, while the minimal coefficient $|\Gamma_{min}|$ is a *local* one – thus making this term enormous, and the corresponding bound nearly meaningless. As we show next, one can harness the inherent locality of the atoms in order to replace the global quantity in the numerator with a local one: ϵ_L.

Theorem 17 (Stable recovery of global OMP in the presence of noise): Suppose a clean signal $\mathbf{X}$ has a representation $\mathbf{D\Gamma}$, and that it is contaminated with noise $\mathbf{E}$ to create the signal $\mathbf{Y} = \mathbf{X} + \mathbf{E}$, such that $\|\mathbf{Y} - \mathbf{X}\|_2 \leq \epsilon$. Denote by ϵ_L the highest energy of all n-dimensional local patches extracted from $\mathbf{E}$. Assume $\mathbf{\Gamma}$ satisfies

$$\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) - \frac{1}{\mu(\mathbf{D})} \cdot \frac{\epsilon_L}{|\Gamma_{min}|},$$

where $|\Gamma_{min}|$ is the minimal entry in absolute value of the sparse vector $\mathbf{\Gamma}$. Denoting by $\mathbf{\Gamma}_{\text{OMP}}$ the solution obtained by running OMP for $\|\mathbf{\Gamma}\|_0$ iterations, we are guaranteed that

a) OMP will find the correct support; And,
b) $\|\mathbf{\Gamma}_{\text{OMP}} - \mathbf{\Gamma}\|_2^2 \leq \frac{\epsilon^2}{1-\mu(\mathbf{D})(\|\mathbf{\Gamma}\|_{0,\infty}-1)}$.

The proof of this theorem is presented in the Supplementary Material, and the derivations therein are based on the analysis presented in [34], generalizing the study to the convolutional setting. Note that we have assumed that the OMP algorithm runs for $\|\mathbf{\Gamma}\|_0$ iterations. We could also propose a different approach, however, using a stopping criterion based on the norm of the residual. Under such setting, the OMP would run until the energy of the global residual is less than the energy of the noise, given by ϵ^2.

C. Stability Guarantee of Basis Pursuit Denoising via ERC

A theoretical motivation behind relaxing the $\ell_{0,\infty}$ norm to the convex ℓ_1 was already established in Section V, showing that if the former is low, the BP algorithm is guaranteed to succeed. When moving to the noisy regime, the BP is naturally extended to the Basis Pursuit DeNoising (BPDN) algorithm,[7] which in its Lagrangian form is defined as follows

$$\min_{\mathbf{\Gamma}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D\Gamma}\|_2^2 + \lambda\|\mathbf{\Gamma}\|_1. \tag{9}$$

Similar to how BP was shown to approximate the solution to the $P_{0,\infty}$ problem, in what follows we will prove that the BPDN manages to approximate the solution to the $P^{\epsilon}_{0,\infty}$ problem.

Assuming the ERC is met, the stability of BP was proven under various noise models and formulations in [36]. By exploiting the convolutional structure used throughout our analysis, we now show that the ERC is met given that the $\ell_{0,\infty}$ norm is small, tying the aforementioned results to our story.

Theorem 18 (ERC in the convolutional sparse model): For a convolutional dictionary $\mathbf{D}$ with mutual coherence $\mu(\mathbf{D})$, the ERC condition is met for every support $\mathcal{T}$ that satisfies

$$\|\mathcal{T}\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

[7]Note that an alternative to the BPDN extension is that of the Dantzig Selector algorithm. One can envision a similar analysis to the one presented here for this algorithm as well.

Based on this and the analysis presented in [36], we present a stability claim for the Lagrangian formulation of the BP problem as stated in Eq. (9).

Theorem 19 (Stable recovery of global Basis Pursuit in the presence of noise): Suppose a clean signal $\mathbf{X}$ has a representation $\mathbf{D}\boldsymbol{\Gamma}$, and that it is contaminated with noise $\mathbf{E}$ to create the signal $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. Denote by ϵ_L the highest energy of all n-dimensional local patches extracted from $\mathbf{E}$. Assume $\boldsymbol{\Gamma}$ satisfies

$$\|\boldsymbol{\Gamma}\|_{0,\infty} \leq \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

Denoting by $\boldsymbol{\Gamma}_{\text{BP}}$ the solution to the Lagrangian BP formulation with parameter $\lambda = 4\epsilon_L$, we are guaranteed that

1) The support of $\boldsymbol{\Gamma}_{\text{BP}}$ is contained in that of $\boldsymbol{\Gamma}$.
2) $\|\boldsymbol{\Gamma}_{\text{BP}} - \boldsymbol{\Gamma}\|_\infty < \frac{15}{2}\epsilon_L$.
3) In particular, the support of $\boldsymbol{\Gamma}_{\text{BP}}$ contains every index i for which $|\Gamma_i| > \frac{15}{2}\epsilon_L$.
4) The minimizer of the problem, $\boldsymbol{\Gamma}_{\text{BP}}$, is unique.

The proof for both of the above, inspired by the derivations in [27] and [36], are presented in the Supplementary Material.

The benefit of this over traditional claims is, once again, the replacement of the ℓ_0 with the $\ell_{0,\infty}$ norm. Moreover, this result bounds the difference between the entries in $\boldsymbol{\Gamma}_{\text{BP}}$ and $\boldsymbol{\Gamma}$ in terms of a local quantity – the local noise level ϵ_L. As a consequence, all atoms with coefficients above this local measure are guaranteed to be recovered.

The implications of the above theorem are far-reaching as it provides a sound theoretical back-bone for all works that have addressed the convolutional BP problem in its Lagrangian form [21]–[23], [37], [46]. In Section IX we will propose two additional algorithms for solving the global BP efficiently by working locally, and these methods would benefit from this theoretical result as well. As a last comment, a different and perhaps more appropriate convex relaxation for the $\ell_{0,\infty}$ norm could be suggested, such as the $\ell_{1,\infty}$ norm. This, however, remains one of our future work challenges.

D. Experiments

Following the above analysis, we now provide a numerical experiment demonstrating the above obtained bounds. The global dictionary employed here is the same as the one used for the noiseless experiments in Section V, with mutual coherence $\mu(\mathbf{D}) = 0.09$, local atoms of length $n = 64$ and global ones of size $N = 640$. We sample random sparse vectors with cardinality between 1 and 500, with entries drawn from a uniform distribution with range $[-a, a]$, for varying values of a. Given these vectors, we construct global signals and contaminate them with noise. The noise is sampled from a zero-mean unit-variance white Gaussian distribution, and then normalized such that $\|\mathbf{E}\|_2 = 0.1$.

In what follows, we will first center our attention on the bounds obtained for the OMP algorithm, and then proceed to the ones corresponding to the BP. Given the noisy signals, we run OMP with a sparsity constraint, obtaining $\boldsymbol{\Gamma}_{\text{OMP}}$. For each realization of the global signal, we compute the minimal entry

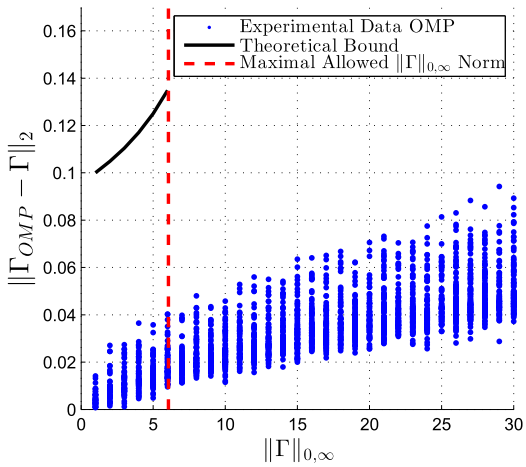


Fig. 4. The distance $\|\boldsymbol{\Gamma}_{\text{OMP}} - \boldsymbol{\Gamma}\|_2$ as a function of the $\ell_{0,\infty}$ norm, and the corresponding theoretical bound.

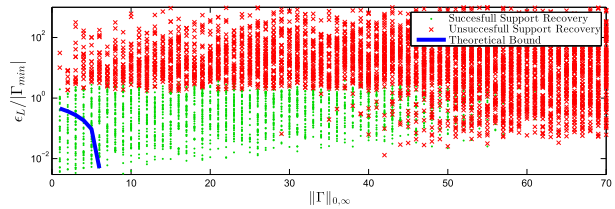
(in absolute value) of the global sparse vector, $|\Gamma_{min}|$, and its $\ell_{0,\infty}$ norm. In addition, we compute the maximal local energy of the noise, ϵ_L, corresponding to the highest energy of a n-dimensional patch of $\mathbf{E}$.

Recall that the theorem in the previous subsection poses two claims: 1) the stability of the result in terms of $\|\boldsymbol{\Gamma}_{\text{OMP}} - \boldsymbol{\Gamma}\|_2$; and 2) the success in recovering the correct support. In Fig. 4 we investigate the first of these points, presenting the distance between the estimated and the true sparse codes as a function of the $\ell_{0,\infty}$ norm of the original vector. As it is clear from the graph, the empirical distances are below the theoretical bound depicted in black, given by $\frac{\epsilon^2}{1-\mu(\mathbf{D})(\|\boldsymbol{\Gamma}\|_{0,\infty}-1)}$. According to the theorem's assumption, the sparse vector should satisfy $\|\boldsymbol{\Gamma}\|_{0,\infty} < \frac{1}{2}(1 + \frac{1}{\mu(\mathbf{D})}) - \frac{1}{\mu(\mathbf{D})} \cdot \frac{\epsilon_L}{|\Gamma_{\min}|}$. The red dashed line delimits the area where this is met, with the exception that we omit the second term in the previous expression, as done previously in [34]. This disregards the condition on the $|\Gamma_{\min}|$ and ϵ_L (which depends on the realization). Yet, the empirical results remain stable.

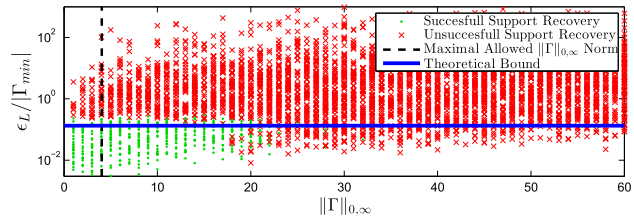
In order to address the successful recovery of the support, we compute the ratio $\frac{\epsilon_L}{|\Gamma_{\min}|}$ for each realization in the experiment. In Fig. 5(a), for each sample we denote by $\bullet$ or $\times$ the success or failure in recovering the support, respectively. Each point is plotted as a function of its $\ell_{0,\infty}$ norm and its corresponding ratio. The theoretical condition for the success of the OMP can be rewritten as $\frac{\epsilon_L}{|\Gamma_{min}|} < \frac{\mu(\mathbf{D})}{2}(1 + \frac{1}{\mu(\mathbf{D})}) - \mu(\mathbf{D})\|\boldsymbol{\Gamma}\|_{0,\infty}$, presenting a bound on the ratio $\frac{\epsilon_L}{|\Gamma_{\min}|}$ as a function of the $\ell_{0,\infty}$ norm. This bound is depicted with a blue line, indicating that the empirical results agree with the theoretical claims.

One can also observe two distinct phase transitions in Fig. 5(a). On the one hand, noting that the y axis can be interpreted as the inverse of the noise-to-signal ratio (in some sense), we see that once the noise level is too high, OMP fails in recovering the support.[8] On the other hand, similar to what was presented in the noiseless case, once the $\ell_{0,\infty}$ norm becomes too large, the algorithm is prone to fail in recovering the support.

[8]Note that the abrupt change in this phase-transition area is due to the log scale of the y axis.



(a) Orthogonal Matching Pursuit.



(b) Basis Pursuit.

Fig. 5. The ratio $\epsilon_L/|\Gamma_{\min}|$ as a function of the $\ell_{0,\infty}$ norm, and the theoretical bound for the successful recovery of the support, for both the OMP and BP algorithms. (a) Orthogonal matching pursuit. (b) Basis pursuit.

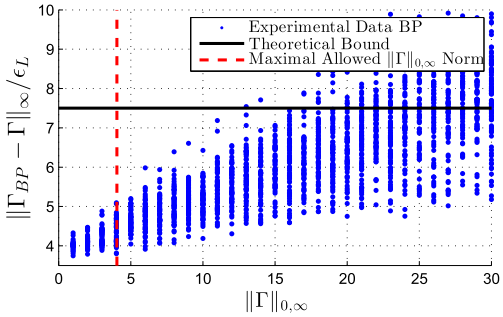


Fig. 6. The distance $\|\mathbf{\Gamma}_{\mathrm{BP}}-\mathbf{\Gamma}\|_\infty/\epsilon_L$ as a function of the $\ell_{0,\infty}$ norm, and the corresponding theoretical bound.

We now shift to the empirical verification of the guarantees obtained for the BP. We employ the same dictionary as in the experiment above, and the signals are constructed in the same manner. We use the implementation of the LARS algorithm within the SPAMS package[9] in its Lagrangian formulation with the theoretically justified parameter $\lambda = 4\epsilon_L$, obtaining $\mathbf{\Gamma}_{\mathrm{BP}}$. Once again, we compute the quantities: $|\Gamma_{min}|$, $\|\mathbf{\Gamma}\|_{0,\infty}$ and ϵ_L.

Theorem 19 states that the ℓ_∞ distance between the BP solution and the true sparse vector is below $\frac{15}{2}\epsilon_L$. In Fig. 6 we depict the ratio $\frac{\|\mathbf{\Gamma}_{\mathrm{BP}}-\mathbf{\Gamma}\|_\infty}{\epsilon_L}$ for each realization, verifying it is indeed below $\frac{15}{2}$ as long as the $\ell_{0,\infty}$ norm is below $\frac{1}{3}(1+\frac{1}{\mu(\mathbf{D})}) \approx 4$. Next, we would like to corroborate the assertions regarding the recovery of the true support. To this end, note that the theorem guarantees that all entries satisfying $|\Gamma_i| > \frac{15}{2}\epsilon_L$ shall be recovered by the BP algorithm. Alternatively, one can state that the complete support must be recovered as long as $\frac{\epsilon_L}{|\Gamma_{\min}|} < \frac{2}{15}$. To verify this claim, we plot this ratio for each realization as function of the $\ell_{0,\infty}$ norm in Fig. 5(b), marking every point according to the success or failure of BP (in recovering the complete support). As evidenced in [27], OMP seems to be far more accurate than the BP in recovering the true support. As one can see by comparing Fig. 5(a) and 5(b), BP fails once the $\ell_{0,\infty}$ norm goes beyond 20, while OMP succeeds all the way until $\|\mathbf{\Gamma}\|_{0,\infty} = 40$.

[9]Freely available from http://spams-devel.gforge.inria.fr/.

IX. From Global Pursuit to Local Processing

We now turn to analyze the practical aspects of solving the $P^{\epsilon}_{0,\infty}$ problem given the relationship $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$. Motivated by the theoretical guarantees of success derived in the previous sections, the first naïve approach would be to employ global pursuit methods such as OMP and BP. However, these are computationally demanding as the dimensions of the convolutional dictionary are prohibitive for high values of N, the signal length.

As an alternative, one could attempt to solve the $P^{\epsilon}_{0,\infty}$ problem using a patch-based processing scheme. In this case, for example, one could suggest to solve a local and relatively cheaper pursuit for every patch in the signal (including overlaps) using the local dictionary $\mathbf{D}_L$. It is clear, however, that this approach will not work well under the convolutional model, because atoms used in overlapping patches are simply not present in $\mathbf{D}_L$. On the other hand, one could turn to employ $\mathbf{\Omega}$ as the *local* dictionary, but this is prone to fail in recovering the correct support of the atoms. To see this more clearly, note that there is no way to distinguish between any of the atoms having only one entry different than zero; i.e., those appearing on the extremes of $\mathbf{\Omega}$ in Fig. 2.

As we can see, neither the naïve global approach, nor the simple patch-based processing, provide an effective strategy. Several questions arise from this discussion: Can we solve the global pursuit problem using local patch-based processing? Can the proposed algorithm rely merely on the low dimensional dictionaries $\mathbf{D}_L$ or $\mathbf{\Omega}$ while still fully solving the global problem? If so, in what form should the local patches communicate in order to achieve a global consensus? In what follows, we address these issues and provide practical and globally optimal answers.

A. Global to Local Through Bi-Level Consensus

When dealing with global problems which can be solved locally, a popular tool of choice is the Alternating Direction Method of Multipliers (ADMM) [26] in its consensus formulation. In this framework, a global objective can be decomposed into a set of local and distributed problems which attempt to reach a global agreement. We will show that this scheme can be effectively applied in the convolutional sparse coding context, providing an algorithm with a bi-level consensus interpretation.

The ADMM has been extensively used throughout the literature in convolutional sparse coding. However, as mentioned in the introduction, it has been usually applied in the Fourier domain. As a result, the sense of locality is lost in these approaches and the connection to traditional (local) sparse coding is non-existent. On the contrary, the pursuit method we propose here is carried out in a localized fashion in the original domain, while still benefiting from the advantages of ADMM.

Recall the ℓ_1 relaxation of the global pursuit, given in Eq. (9). Note that the noiseless model is contained in this formulation as

Algorithm 1: Locally operating global pursuit via ADMM.

while *not converged* **do**

Local Thresholding:
$\boldsymbol{\alpha}_i \leftarrow \min_{\boldsymbol{\alpha}} \lambda\|\boldsymbol{\alpha}\|_1 + \frac{\rho}{2}\|\mathbf{Q}\boldsymbol{\gamma}_i - \boldsymbol{\alpha} + \mathbf{u}_i\|_2^2$;

Stripe Projection:

$$\boldsymbol{\gamma}_i \leftarrow \mathbf{Z}^{-1}\left(\frac{1}{n}\boldsymbol{\Omega}^T\mathbf{R}_i\mathbf{Y} + \rho(\mathbf{S}_i\boldsymbol{\Gamma} + \bar{\mathbf{u}}_i) + \rho\mathbf{Q}^T(\boldsymbol{\alpha}_i - \mathbf{u}_i)\right),$$

where $\mathbf{Z} = \rho\mathbf{Q}^T\mathbf{Q} + \frac{1}{n}\boldsymbol{\Omega}^T\boldsymbol{\Omega} + \rho\mathbf{I}$;

Global Update:
$\boldsymbol{\Gamma} \leftarrow \left(\sum_i \mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\sum_i \mathbf{S}_i^T(\boldsymbol{\gamma}_i - \bar{\mathbf{u}}_i)$;

Dual Variables Update:
$\mathbf{u}_i \leftarrow \mathbf{u}_i + (\mathbf{Q}\boldsymbol{\gamma}_i - \boldsymbol{\alpha}_i)$;
$\bar{\mathbf{u}}_i \leftarrow \bar{\mathbf{u}}_i + (\mathbf{S}_i\boldsymbol{\Gamma} - \boldsymbol{\gamma}_i)$;

end

a particular case when λ tends to zero. Using the separability of the ℓ_1 norm, $\|\boldsymbol{\Gamma}\|_1 = \sum_i \|\boldsymbol{\alpha}_i\|_1$, where $\boldsymbol{\alpha}_i$ are $m-$dimensional local sparse vectors, as previously defined. In addition, using the fact that $\mathbf{R}_i\mathbf{D}\boldsymbol{\Gamma} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i$, we apply a local decomposition on the first term as well. This results in

$$\min_{\{\boldsymbol{\alpha}_i\},\{\boldsymbol{\gamma}_i\}} \quad \frac{1}{2n}\sum_i \|\mathbf{R}_i\mathbf{Y} - \boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2^2 + \lambda\sum_i \|\boldsymbol{\alpha}_i\|_1,$$

where we have divided the first sum by the number of contributions per entry in the global signal, which is equal to the patch size n. Note that the above minimization is not equivalent to the original problem in Eq. (9) since no explicit consensus is enforced between the local variables. Recall that the different $\boldsymbol{\gamma}_i$ overlap, and so we must enforce them to agree. In addition, $\boldsymbol{\alpha}_i$ should be constrained to be equal to the center of the corresponding $\boldsymbol{\gamma}_i$. Based on these observations, we modify the above problem by adding the appropriate constraints, obtaining

$$\min_{\{\boldsymbol{\alpha}_i\},\{\boldsymbol{\gamma}_i\},\boldsymbol{\Gamma}} \quad \frac{1}{2n}\sum_i \|\mathbf{R}_i\mathbf{Y} - \boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2^2 + \lambda\sum_i \|\boldsymbol{\alpha}_i\|_1 \quad \text{s.t.} \begin{cases} \mathbf{Q}\boldsymbol{\gamma}_i = \boldsymbol{\alpha}_i \\ \mathbf{S}_i\boldsymbol{\Gamma} = \boldsymbol{\gamma}_i \end{cases} \quad \forall i,$$

where $\mathbf{Q}$ extracts the center m coefficients corresponding to $\boldsymbol{\alpha}_i$ from $\boldsymbol{\gamma}_i$, and $\mathbf{S}_i$ extracts the ith stripe $\boldsymbol{\gamma}_i$ from $\boldsymbol{\Gamma}$.

Defining $f_i(\boldsymbol{\gamma}_i) = \frac{1}{2n}\|\mathbf{R}_i\mathbf{Y} - \boldsymbol{\Omega}\boldsymbol{\gamma}_i\|_2^2$ and $g(\boldsymbol{\alpha}_i) = \lambda\|\boldsymbol{\alpha}_i\|_1$, the above problem can be minimized by employing the ADMM algorithm, as depicted in Algorithm 1. This is a two-level local-global consensus formulation: each m dimensional vector $\boldsymbol{\alpha}_i$ is enforced to agree with the center of its corresponding $(2n-1)m$ dimensional $\boldsymbol{\gamma}_i$, and in addition, all $\boldsymbol{\gamma}_i$ are required to agree with each other as to create a global $\boldsymbol{\Gamma}$. The above can be shown to be equivalent to the standard two-block ADMM formulation [26]. Each iteration of this method can be divided into four steps:

1) Local sparse coding that updates $\boldsymbol{\alpha}_i$ (for all i), which amounts to a simple soft thresholding operation.
2) Solution of a linear system of equations for updating $\boldsymbol{\gamma}_i$ (for all i), which boils down to a simple multiplication by a constant matrix.
3) Update of the global sparse vector $\boldsymbol{\Gamma}$, which aggregates the $\boldsymbol{\gamma}_i$ by averaging.
4) Update of the dual variables.

As can be seen, the ADMM provides a simple way of breaking the global pursuit into local operations. Moreover, the local coding step is just a projection problem onto the ℓ_1 ball, which can be solved through simple soft thresholding, implying that there is no complex pursuit involved.

Since we are in the ℓ_1 case, the function g is convex, as are the functions f_i. Therefore, the above is guaranteed to converge to the minimizer of the global BP problem. As a result, we benefit from the theoretical guarantees derived in previous sections. One could attempt, in addition, to enforce an ℓ_0 penalty instead of the ℓ_1 norm on the global sparse vector. Despite the fact that no convergence guarantees could be claimed under such formulation, the derivation of the algorithm remains practically the same, with the only exception that the soft thresholding is replaced by a hard one.

B. An Iterative Soft Thresholding Approach

While the above algorithm suggests a way to tackle the global problem in a local fashion, the matrix involved in the stripe projection stage, $\mathbf{Z}^{-1}$, is relatively large when compared to the dimensions of $\mathbf{D}_L$. As a consequence, the bi-level consensus introduces an extra layer of complexity to the algorithm. In what follows, we propose an alternative method based on the Iterative Soft Thresholding (IST) algorithm that relies solely on multiplications by $\mathbf{D}_L$ and features a simple intuitive interpretation and implementation. A similar approach for solving the convolutional sparse coding problem was suggested in [47]. Our main concern here is to provide insights into local alternatives for the global sparse coding problem and their guarantees, whereas the work in [47] focused on the optimizations aspects of this pursuit from an entirely global perspective.

Let us consider the IST algorithm [48] which minimizes the global objective in Eq. (9), by iterating the following updates

$$\boldsymbol{\Gamma}^k = \mathcal{S}_{\lambda/c}\left(\boldsymbol{\Gamma}^{k-1} + \frac{1}{c}\mathbf{D}^T(\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}^{k-1})\right),$$

where $\mathcal{S}$ applies an entry-wise soft thresholding operation with threshold λ/c. Interpreting the above as a projected gradient descent, the coefficient c relates to the gradient step size and should be set according to the maximal singular value of the matrix $\mathbf{D}$ in order to guarantee convergence [48].

The above algorithm might at first seem undesirable due to the multiplications of the residual $\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}^{k-1}$ with the global dictionary $\mathbf{D}$. Yet, as we show in the Supplementary Material, such a multiplication does not need to be carried out explicitly due to the convolutional structure imposed on our dictionary. In fact, the above is mathematical equivalent to an algorithm that performs local updates given by

$$\boldsymbol{\alpha}_i^k = \mathcal{S}_{\lambda/c}\left(\boldsymbol{\alpha}_i^{k-1} + \frac{1}{c}\,\mathbf{D}_L^T\,\mathbf{r}_i^{k-1}\right),$$

Algorithm 2: Global pursuit using local processing via iterative soft thresholding.

$\forall i \quad \mathbf{r}_i^0 = \mathbf{R}_i\mathbf{Y}, \quad \boldsymbol{\alpha}_i^0 = \mathbf{0}$;
k = 1;
while *not converged* **do**

Local Coding:
$\forall i \quad \boldsymbol{\alpha}_i^k = \mathcal{S}_{\lambda/c}\left(\boldsymbol{\alpha}_i^{k-1} + \frac{1}{c}\,\mathbf{D}_L^T\,\mathbf{r}_i^{k-1}\right)$;

Computation of the Patch Averaging Aggregation:
$\widehat{\mathbf{X}}^k = \sum_i \mathbf{R}_i^T\mathbf{D}_L\boldsymbol{\alpha}_i^k$;

Update of the Residuals:
$\forall i \quad \mathbf{r}_i^k = \mathbf{R}_i\left(\mathbf{Y} - \widehat{\mathbf{X}}^k\right)$;

$k = k + 1$;
end

where $\mathbf{r}_i^k = \mathbf{R}_i(\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}^{k-1})$ is a patch from the global residual. This scheme is depicted in Algorithm 2.

From an optimization point of view, one can interpret each iteration of the above as a scatter and gather process: local residuals are first extracted and scattered to different nodes where they undergo shrinkage operations, and the results are then gathered for the re-computation of the global residual. From an image processing point of view, this algorithm decomposes a signal into overlapping patches, *restores* these separately and then aggregates the result for the next iteration. Notably, this is very reminiscent of the patch averaging scheme, as described in the introduction, and it shows for the first time the relation between patch averaging and the convolutional sparse model. While the former processes every patch once and independently, the above algorithm indicates that one must iterate this process if one is to reach global consensus.

Assuming the step size is chosen appropriately, the above algorithm is also guaranteed to converge to the solution of the global BP. As such, our theoretical analysis holds in this case as well. Alternatively, one could attempt to employ an ℓ_0 approach, using a global iterative hard thresholding algorithm. In this case, however, there are no theoretical guarantees in terms of the $\ell_{0,\infty}$ norm. Still, we believe that a similar analysis to the one taken throughout this work could lead to such claims.

C. Experiments

Next, we proceed to provide empirical results for the above described methods. To this end, we take an undercomplete DCT dictionary of size 25×5, and use it as $\mathbf{D}_L$ in order to construct the global convolutional dictionary $\mathbf{D}$ for a signal of length $N = 300$. We then generate a random global sparse vector $\boldsymbol{\Gamma}$ with 50 non-zeros, with entries distributed uniformally in the range $[-2,-1] \cup [1,2]$, creating the signal $\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$.

We first employ the ADMM and IST algorithms in a noiseless scenario in order to minimize the global BP and find the underlying sparse vector. Since there is no noise added in this case, we decrease the penalty parameter λ progressively throughout the iterations, making this value tend to zero as suggested in the previous subsection. In Fig. 7 we present the evolution of the estimated $\hat{\boldsymbol{\Gamma}}$ for the ADMM solver throughout the iterations, after the global update stage. Note how the algorithm progressively increases the consensus and eventually recovers the true sparse vector. Equivalent plots are obtained for the IST method, and these are therefore omitted.

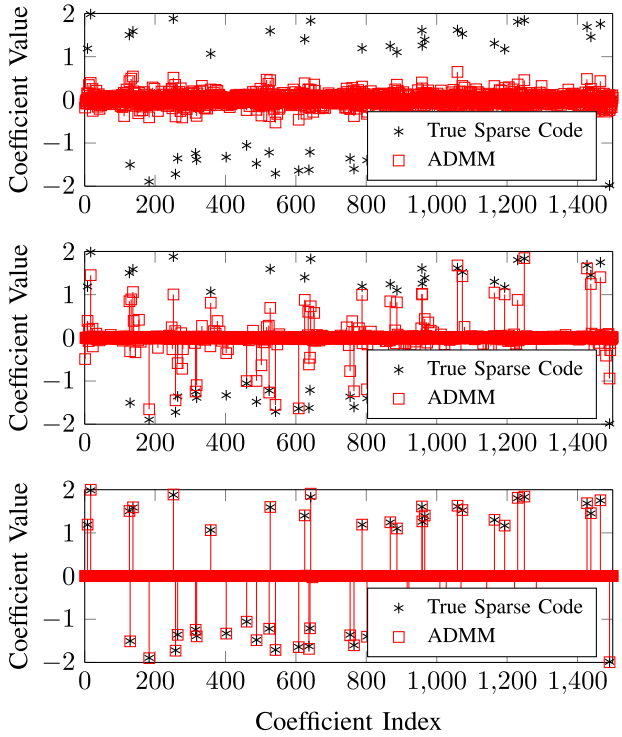


Fig. 7. The sparse vector $\boldsymbol{\Gamma}$ after the global update stage in the ADMM algorithm at iterations 20 (top), 200 (middle) and 1000 (bottom). An ℓ_1 norm formulation was used for this experiment, in a noiseless setting.

To extend the experiment to the noisy case, we contaminate the previous signal with additive white Gaussian noise of different standard deviations: $\sigma = 0.02, 0.04, 0.06$. We then employ both local algorithms to solve the corresponding BPDN problems, and analyze the ℓ_2 distance between their estimated sparse vector and the true one, as a function of time. These results are depicted in Fig. 8, where we include for completion the distance of the solution achieved by the global BP in the noisy cases. A few observations can be drawn from these results. Note that both algorithms converge to the solution of the global BP in all cases. In particular, the IST converges significantly faster than the ADMM method. Interestingly, despite the later requiring a smaller number of iterations to converge, these are relatively more expensive than those of the IST, which employs only multiplications by the small $\mathbf{D}_L$.

X. Conclusion and Future Work

In this work we have presented a formal analysis of the convolutional sparse representation model. In doing so, we have reformulated the objective of the global pursuit, introducing the $\ell_{0,\infty}$ norm and the corresponding $P_{0,\infty}$ problem, and proven the uniqueness of its solution. By migrating from the P_0 to the $P_{0,\infty}$ problem, we were able to provide meaningful guarantees for the

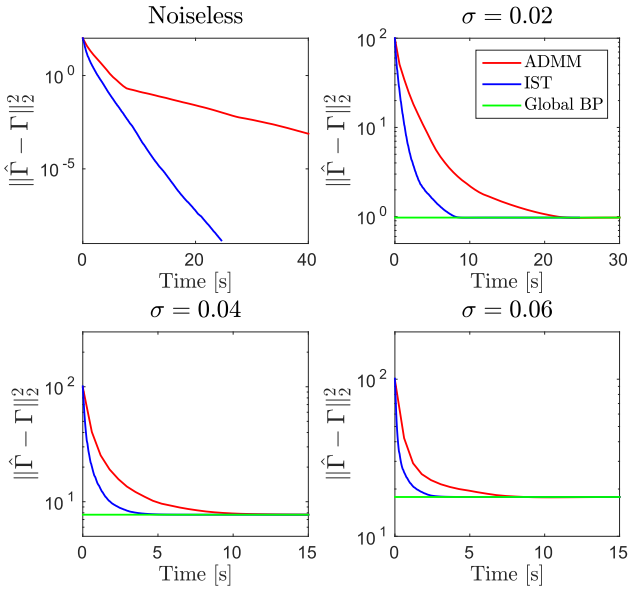


Fig. 8. Distance between the estimate $\hat{\mathbf{\Gamma}}$ and the underlying solution $\mathbf{\Gamma}$ as a function of time for the IST and the ADMM algorithms compared to the solution obtained by solving the global BP.

success of popular algorithms in the noiseless case, improving on traditional bounds that were shown to be very pessimistic under the convolutional case. In order to achieve such results, we have generalized a series of concepts such as Spark and the mutual coherence to their counterparts in the convolutional setting.

Striding on the foundations paved in the first part of this work, we moved on to present a series of stability results for the convolutional sparse model in the presence of noise, providing guarantees for corresponding pursuit algorithms. These were possible due to our migration from the ℓ_0 to the $\ell_{0,\infty}$ norm, together with the generalization and utilization of concepts such as RIP and ERC. Seeking for a connection between traditional patch-based processing and the convolutional sparse model, we finally proposed two efficient methods that solve the global pursuit while working locally.

We envision many possible directions of future work, and here we outline some of them:

- We could extend our study, which considers only worst-case scenarios, to an average-performance analysis. By assuming more information about the model, it might be possible to quantify the probability of success of pursuit methods in the convolutional case. Such results would close the gap between current bounds and empirical results.
- From an application point of view, we envision that interesting algorithms could be proposed to tackle real problems in signal and image processing while using the convolutional model. We note that while convolutional sparse coding has been applied to various problems, simple inverse problems such as denoising have not yet been properly addressed. We believe that the analysis presented in this work could facilitate the development of such algorithms by showing how to leverage on the subtleties of this model.
- Interestingly, even though we have declared the $P_{0,\infty}$ problem as our goal, at no point have we actually attempted to tackle it directly. What we have shown instead is that popular algorithms succeed in finding its solution. One could perhaps propose an algorithm specifically tailored for solving this problem – or its convex relaxation ($\ell_{1,\infty}$). Such a method might be beneficial from both a theoretical and a practical aspect.

All these points, and more, are matter of current research.

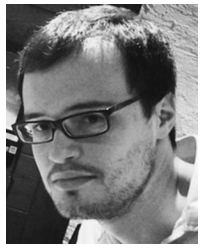
ACKNOWLEDGMENT

The authors would like to thank Dmitry Batenkov, Yaniv Romano and Raja Giryes for the prolific conversations and most useful advice which helped shape this work.

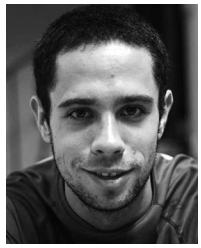
REFERENCES

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.
[2] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations Trends® Comput. Graph. Vis.*, vol. 8, no. 2-3, pp. 85–283, 2014.
[3] Y. Romano, M. Protter, and M. Elad, "Single image interpolation via adaptive nonlocal sparsity-based modeling," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3085–3098, Jul. 2014.
[4] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.
[5] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[6] M. Elad, J. Starck, P. Querre, and D. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 3, pp. 340–358, Nov. 2005.
[7] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 1999, pp. 2443–2446.
[8] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 689–696.
[10] X. Li, "Image recovery via hybrid sparse representations: A deterministic annealing approach," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 953–962, Sep. 2011.
[11] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2691–2698.
[12] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3194–3205, Jul. 2012.
[13] M. Yang, D. Dai, L. Shen, and L. Gool, "Latent dictionary learning for sparse representation based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4138–4145.
[14] J. Sulam and M. Elad, "Expected patch log likelihood with a sparse prior," in *Energy Minimization Methods in Computer Vision and Pattern Recognition* (Lecture Notes in Computer Science). New York, NY, USA: Springer, 2015, pp. 99–111. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14612-6_8
[15] Y. Romano and M. Elad, "Patch-disagreement as away to improve K-SVD denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 1280–1284.
[16] Y. Romano and M. Elad, "Boosting of image denoising algorithms," *SIAM J. Imag. Sci.*, vol. 8, no. 2, pp. 1187–1219, 2015.
[17] V. Papyan and M. Elad, "Multi-scale patch-based image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 249–261, Jan. 2016.

[18] D. Batenkov, Y. Romano, and M. Elad, "On the global-local dichotomy in sparsity modeling," 2017, arXiv:1702.03446.
[19] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 479–486.
[20] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-Invariant sparse coding for audio classification," in *Proc. Conf. Uncertainty Artif. Intell.*, 2007, pp. 149–158.
[21] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 391–398. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6618901
[22] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5135–5143.
[23] H. Bristow, A. Eriksson and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 391–398.
[24] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7173–7177.
[25] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1823–1831.
[26] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
[27] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. New York, NY, USA: Springer, 2010.
[28] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[29] J. Tropp, "Greed is Good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[30] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.
[31] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 1993, pp. 40–44.
[32] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, 1989.
[33] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.
[34] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[35] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.
[36] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[37] H. Bristow and S. Lucey, "Optimization methods for convolutional sparse coding," arXiv preprint arXiv:1406.2407, 2014.
[38] M. Mørup, M. N. Schmidt, and L. K. Hansen, "Shift invariant sparse coding of image and music data," DTU Informatics Technical Univ. of Denmark, Lyngby, Denmark. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=4659
[39] Y. Zhu and S. Lucey, "Convolutional sparse coding for trajectory reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 529–540, Mar. 2015.
[40] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2528–2535.
[41] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1090–1098.
[42] F. Huang and A. Anandkumar, "Convolutional dictionary learning through tensor factorization," in *Feature Extraction: Modern Questions and Challenge*, pp. 116–129, 2015.
[43] L. R. Welch, "Lower bounds on the maximum cross correlation of signals (corresp.)," *IEEE Trans. Inf. Theory*, vol. 20, no. 3, pp. 397–399, May 1974.
[44] M. Soltanalian, M. M. Naghsh, and P. Stoica, "Approaching peak correlation bounds via alternating projections," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 5317–5312.
[45] H. He, P. Stoica, and J. Li, "Designing unimodular sequence sets with good correlations—Including an application to MIMO radar," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4391–4405, Nov. 2009.
[46] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.
[47] R. Chalasani, J. C. Principe, and N. Ramakrishnan, "A fast proximal method for convolutional sparse coding," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–5.
[48] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.



Vardan Papyan received the B.Sc. degree in 2013 from the Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel, where he is currently working toward the Ph.D. degree. His research interests include signal and image processing, sparsity-based modeling of signals, and in particular deep learning and its relation to sparsity.



Jeremias Sulam (S'14) received the Bioengineering degree from the Faculty of Engineering, Universidad Nacional de Entre Rios, Argentina, in 2013. He is currently working toward the Ph.D. degree in the Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel. His research interests include signal and image processing, sparse representation modelling and its application to inverse problems and machine learning.

Michael Elad (F'12) received the B.Sc., M.Sc., and D.Sc. degrees from the Department of Electrical engineering, Technion, Israel, in 1986, 1988, and 1997, respectively. Since 2003, he has been a faculty member in the Computer-Science Department, Technion, and since 2010, he holds a Full-Professorship Position. He works in the field of signal and image processing, specializing in inverse problems, and sparse representations. He received numerous teaching awards, and also the 2008 and 2015 Henri Taub Prizes for Academic Excellence, and the 2010 Hershel-Rich prize for innovation. He is working as the Editor-in-Chief for *SIAM Journal on Imaging Sciences* since January 2016.